JOURNAL OF MULTIVARIATE ANALYSIS 26, 1–15 (1988)

Parametric Estimation for the Mean of a Gaussian Process by the Method of Sieves

ANESTIS ANTONIADIS*

University of California at Irvine

In this paper, we apply Grenader's method of sieves to the problem of estimation of the infinite dimensional mean parameter of a Gaussian random vector with values in a real separable Banach space. We use an increasing sequence of natural sieves in terms of geometric properties of the parameter space, on which we maximize the likelihood function. Sufficient conditions on the growth of the sieves are given in order that the sequence of restricted maximum likelihood estimators of the unknown mean is consistent when the sample size tends to infinity. Exponential rates of convergence in the topology of the parameter space are obtained. Applications to some examples are also discussed. © 1988 Academic Press, Inc.

1. INTRODUCTION

Banach space-valued random vectors arise naturally in the analysis of stochastic processes whose sample paths belong to a function space. Statistical inference for processes with a Gaussian distribution has been investigated by several authors (Antoniadis [2], Beder [7], Geman and Hwang [11], Grenander [12]) often by extending the tools of multivariate analysis. This paper discusses the problem of estimating the mean of a Gaussian random vector with values in a real separable Banach space when a sample of i.i.d. observations is available and the parameter lies in some closed subset, not necessarilly a subspace, of the topological space of means.

In the finite dimensional parameters case, the maximum likelihood principle is widely and successfully used as a method of estimation. In the case of infinite dimensional parameters, it is known that the maximum

Received April 1985; revised May 26, 1987.

AMS 1980 subject classifications: Primary 62M09; Secondary 62A10, 60G15, 62G05.

Key words and phrases: inference for stochastic processes, method of sieves, maximum likelihood estimation, Gaussian process, GB and GC sets, reproducing kernel Hilbert space.

* On leave of absence from Department of Mathematics, Université de Saint-Etienne, 23, rue du Docteur Paul Michelon, 42100 Saint-Etienne, France.

0047-259X/88 $3.00

likelihood method often fails because, either the maximum likelihood solution is not attained or is not consistent. For such problems, another approach is described by U. Grenander in [12], where he suggests performing the maximization of the likelihood function within a constrained subset of the parameter set and then allowing this subset to "grow" with the sample size. He calls this collection of subsets, from which a sequence of estimators is drawn, a *sieve*, and the resulting estimation procedure is his *method of sieves*. A recent survey on applications of the method of sieves to various inference problems is given by McKeague [16].

In the finite dimensional case, when one can observe an i.i.d. sample of Gaussian vectors in $\mathbf{B}$, the sample mean is the maximum likelihood estimator of the mean and it is almost surely consistent. When the dimension of $\mathbf{B}$ is infinite, the sample mean is no longer the maximum likelihood estimator of the mean, but a strong law of large numbers still holds when $\mathbf{B}$ is a real separable Banach space, and the sample mean still defines a sequence of strongly consistent estimators of the expectation parameter, with exponential rates of convergence. But consistency is, in this case, given in terms of the topology of $\mathbf{B}$, instead of the stronger Hilbertian topology of the space of expectations which is the natural topology to consider in this case, by the way the probability distributions depend upon their expectation parameter. The main purpose of this work is to show that in such a situation, a sieve can be chosen simply in order that the restricted maximum likelihood estimators exist and are consistent in the topology of the parameter space.

More precisely, let $\mathbf{B}$ be a real separable Banach space on whose sigma algebra $\mathscr{B}$ of Borel subsets is defined a zero-mean Gaussian measure μ with support $\mathbf{B}$ and covariance kernel K_μ on $\mathbf{B}^* \times \mathbf{B}^*$, where $\mathbf{B}^*$ denotes the topological dual of $\mathbf{B}$. The reproducing kernel Hilbert space (RKHS for short) $H(K_\mu)$ of μ can be realized as a subset of $\mathbf{B}$ and is isometric to the closure of $\mathbf{B}^*$ in $L^2(\mathbf{B}, \mu)$. If μ_θ denotes the Gaussian measure on B with expectation θ and covariance kernel K_μ, the set of θ's for which the measures μ_θ and μ are mutually absolutely continuous is the space $H(K_\mu)$. For any θ in $H(K_\mu)$, the density of μ_θ with respect to μ is given by

$$\frac{d\mu_\theta}{d\mu}(x) = f(x;\theta) = \exp\left[\tilde{\theta}(x) - \frac{1}{2}\|\theta\|_\mu^2\right] \tag{1.1}$$

defined on a set of μ-measure 1, where $\theta \to \tilde{\theta}$ denotes the isometry of $H(K_\mu)$ into $L^2(\mathbf{B}, \mu)$. Here and in the following, $\|\cdot\|_\mu$ denotes the norm of $H(K_\mu)$.

Traditional maximum likelihood estimation consists in finding, for a fixed observation x, the value of θ that maximizes $f(x;\theta)$. Such an optimization is impossible for Eq. (1.1) without further restrictions on θ. Observe first that $\tilde{\theta}(x)$ cannot be defined in a natural way for all θ in

$H(K_\mu)$ simultaneously. Even in the particular case of θ running through $\mathbf{B}^*$, we have (see [4, 6]):

$$\operatorname*{Sup}_{\theta \in \mathbf{B}^*} \left[\tilde{\theta}(x) - \frac{1}{2} \|\theta\|_\mu^2 \right] = \begin{cases} \frac{1}{2}\|x\|_\mu^2 & \text{if} \quad x \in H(K_\mu) \\ \infty & \text{if not} \end{cases}$$

and (1.2)

$$\tilde{\theta}(x) = \langle \theta, x \rangle$$

Since μ-almost every x does not lie in $H(K_\mu)$ (e.g., [19]), the above expression is infinite and the maximum likelihood method fails. The difficulty is due to the fact that the parameter space is too "large." We have to restrict it and the method of sieves does that. A sieve, generally speaking, is a family of subsets $\{\Theta_\lambda; \lambda > 0\}$ of the parameter set Θ. For any $\lambda > 0$, Θ_λ is sufficiently restricted to make maximum likelihood solutions exists. On the other hand, the sets Θ_λ are chosen to be sufficiently rich so that, as λ tends toward zero and the sample size tends to infinity, maximum likelihood solutions converge to the value of the true parameter.

The results of the following sections are motivated by some of the work of Geman and Hwang [11] where one can find a general theorem declaring the existence and consistency of maximum likelihood estimators. Their theorem is not specific to the Gaussian case and leads to a quite complicate recipe for computing the rates of growth of the sieve size. Most of the work is left to the application of rather complicated conditions to specific examples. Our results are simpler because we consider only Gaussian measures. They are stronger in the sense that they provide rates for the speed of convergence of the estimators to the true value of the unknown parameter.

The paper is organized in order to make it understandable to the reader who is not quite familiar with either the theory of the method of sieves or the theory of Banach space valued Gaussian variables. Section 2 contains some preliminary material on Gaussian measures on Banach spaces and some notation that are used throughout this work. In Section 3, the solution to the estimation problem of the mean θ, when it belongs to any closed subset Θ of $H(K_\mu)$, is developed. Finally, Section 4 discusses some examples and Section 5 presents some concluding remarks.

2. PRELIMINARIES

2.1. *Notation and Definitions*

For any Banach space $\mathbf{B}$, we will denote by $\|\cdot\|_{\mathbf{B}}$ its norm, by $U_{\mathbf{B}}$ the closed unit ball of $\mathbf{B}$ and by $\langle \cdot, \cdot \rangle$ the pairing $(\mathbf{B}^*, \mathbf{B})$ between the

topological dual $\mathbf{B}^*$ of $\mathbf{B}$ and $\mathbf{B}$. As usual, $L^0(\Omega, \mathcal{A}, P)$ denotes the space of all (classes of P-equivalent) random variables on a given probability space $(\Omega, \mathcal{A}, P)$. Given a random function $X: T \to L^0$ on a set T (resp. on a topological space T), a subset C of T is called a set of boundedness for X (resp. a set of continuity for X) if there exists a modification of X with bounded (resp. continuous) sample functions on C (see, for example, [5, 8, or 9]).

If $\mathbf{B}$ is a real separable Banach space and μ is a mean zero probability measure on the σ-field of Borel subsets of $\mathbf{B}$ such that $\int_{\mathbf{B}} \|x\|_{\mathbf{B}}^2 \mu(dx) < \infty$, then the bilinear functional K_μ defined on $\mathbf{B}^* \times \mathbf{B}^*$ by

$$K_\mu(f, g) = \int_B \langle f, g \rangle \langle g, x \rangle \mu(dx)$$

is called the covariance function of μ.

Let $\{P_\theta; \theta \in \Theta\}$ be a family of probability measures on $(\Omega, \mathcal{A}, P)$ which are absolutely continuous with respect to P, and define $f(\omega; \theta) = dP_\theta/dP(\omega)$. If Θ is a metric space with metric d, we will use the following notation and definitions in the same spirit as that of [11]:

For any extended real function g on Θ and any subset S of Θ, $g(S) = \mathrm{Sup}_{\theta \in S} g(\theta)$. For any couple (θ, θ') in $\Theta \times \Theta$, the formal entropy of P_θ w.r.t. $P_{\theta'}$ is defined by

$$H(\theta, \theta') = E_\theta[\ln f(\cdot, \theta')] = \int_\Omega \ln f(\omega, \theta') P_\theta(d\omega). \tag{2.1}$$

$H(\theta, \theta) - H(\theta, \theta')$ is the familiar Kullback–Leibler information number and represents, in a statistical sense, the average information per observation, when sampling from P_θ, for discrimination against $P_{\theta'}$.

For any problem of inference in the statistical space $(\Omega, \mathcal{A}, \{P_\theta; \theta \in \Theta\})$, θ_0 will denote the true value of the parameter and $P_\theta^{(n)}$ will denote the product measure $P_\theta \times P_\theta \times \cdots \times P_\theta$ on $(\Omega^n, \mathcal{A}^n)$ corresponding to an i.i.d. sample of size n from $(\Omega, \mathcal{A}, P_\theta)$. A sequence $\{\theta_n, n \in \mathbf{N}\}$ of random variables on $(\Omega, \mathcal{A})$ is said to be consistent for a given metric d on Θ if $d(\theta_n, \theta_0) \to 0$ with $P_{\theta_0}^{(\infty)}$-probability one, as n converges to ∞. Finally if Θ_λ is a given subset of Θ, the set of maximum likelihood solutions in Θ_λ, given a sample of size n, will be denoted $M_\lambda(\omega_1, ..., \omega_n)$, and $A_\lambda(\theta_0) = \{\theta \in \Theta_\lambda; H(\theta_0; \theta) = H(\theta_0, \Theta_\lambda)\}$ will be the maximum entropy set in Θ_λ.

2.2. *Gaussian Measures on Banach Spaces*

Most of the facts which we state here are well known and can be found in [5] and [9]. Another good source is [15]. For the convenience of the reader we have reproduced part of them in this subsection.

If μ is a mean zero Gaussian measure on a real separable Banach space **B**, it is known that the reproducing kernel Hilbert space $H(K_\mu)$ of μ can be realized as a measurable dense subset of **B**, and μ appears as the σ-additive extension of the canonical cylinder Gaussian measure on $H(K_\mu)$. If i denotes the canonical embedding of $H(K_\mu)$ into **B**, the triple $(i, H(K_\mu), B)$ is an abstract Wiener space [13]. By duality, $\mathbf{B}^*$ is embedded, by i^*, with a dense image into $H^*(K_\mu)$ and we shall identify $H^*(K_\mu)$ and $H(K_\mu)$ via the Riesz identification. Elements of $\mathbf{B}^*$ interpreted as functions on **B** belong to $L^2(\mathbf{B}, \mathscr{B}, \mu)$ and their L^2-norm is equal to their $H^*(K_\mu)$-norm; hence, the closure H^c of $\mathbf{B}^*$ in $L^2(\mathbf{B}, \mathscr{B}, \mu)$ is isomorphic to $H(K_\mu)$. If L denotes the isometry of $H(K_\mu)$ into H^c, the Gaussian random function L: $H(K_\mu) \to L^0(\mathbf{B}, \mathscr{B}, \mu)$ is called the isonormal process on $H(K_\mu)$ associated to μ.

A subset C of $H(K_\mu)$ is called a GB-set (resp. a GC-set) iff C is a set of boundedness (resp. of continuity) for the isonormal process L on $H(K_\mu)$. The closure of a GB-set in $\mathrm{H}(\mathrm{K}_\mu)$ is a compact GB-set and Feldman [10] proved that any bounded convex symmetric GC-set of $H(K_\mu)$ is a GB-set and that the closure of such a set is also a GC-set.

An important numerical measure of the size of a subset C of $H(K_\mu)$ is the mean width of C. It arises naturally in the study of the GB and GC properties of a set C. More precisely, for a subset C of $H(K_\mu)$, the mean width of C is given by the number

$$h(C) = (2\pi)^{1/2} \int_B (\sup_{\theta \in C} L(\theta)(y))\, \mu(dy). \tag{2.2}$$

On $\wp(H(K_\mu)) - \{\varnothing\}$, h is positive, increasing with respect to inclusion, homothetically invariant, lower semicontinuous, and additive with respect to algebraic addition (i.e., $h(K) + h(K') = h(K + K')$) (see [5 or 18]). Moreover, we have:

THEOREM 2.1 [5, 18]. *Let A be a nonempty bounded set of $H(K_\mu)$. The set A is a GB-set iff $h(A) < \infty$. If A is convex and symmetric (resp. compact) then A is a GC-set iff for any (resp. one) increasing sequence $(\Pi_n)_{n \in \mathbf{N}}$ of finite dimensional projections of $H(K_\mu)$ converging strongly to the identity operator of $H(K_\mu)$, $h(\Pi_n^\perp(A))$ tends to zero.*

We end this section by recalling that if θ is in $H(K_\mu)$ and μ_θ is the Gaussian measure on $(\mathbf{B}, \mathscr{B})$ with mean θ and covariance function K_μ, then μ_θ and μ are equivalent and the *R-N* derivative of μ_θ with respect to μ is given by

$$\frac{d\mu_\theta}{d\mu} = \exp\left\{L(\theta) - \frac{1}{2}\|\theta\|_\mu^2\right\}. \tag{2.3}$$

3. Maximum Likelihood Sieve Estimation

From now on **B** will be a real separable Banach space and μ a mean zero Gaussian probability measure on $\mathscr{B}$ whose covariance function K_μ is assumed to be known. This measure accomplishes the task of a carrier measure for the particular statistical space that we will analyze. According to the notation of Section 2, we consider here the problem of estimation of the unknown mean θ_0 connected with the observation of an i.i.d. sample in the statistical space

$$[\mathbf{B}, \mathscr{B}, \{\mu_\theta; \theta \in \Theta\}]$$

where Θ is a closed subset of the RKHS $H(K_\mu)$ of μ. It is usefull to call Θ the natural parameter set.

A sieve for the parameter set Θ is an increasing sequence $(\Theta_n)_{n \in \mathbf{N}}$ of compact subsets of Θ such that $\bigcup_{n \in \mathbf{N}} \Theta_n$ is dense in Θ.

To guarantee the asymptotic consistency in the norm of $H(K_\mu)$ of the maximum likelihood sieve estimators, a natural sieve and an appropriate sieve size will be chosen. Moreover, this choice will allow us to prove that the maximum likelihood sets have only one element, thus avoiding problems of measurability. Since the empirical mean is a sufficient statistic for θ, we shall estimate the parameter θ_0 on the basis of a single observation x in **B**, if by x we mean the average of the sample; the distribution of this statistic is the Gaussian measure on $\mathscr{B}$, homothetic to the initial one and corresponds to the statistical space

$$[\mathbf{B}, \mathscr{B}, \{\mu_\theta^n; \theta \in \Theta\}], \tag{3.1}$$

where μ_θ^n is the image of the measure μ under the map $x \to (1/\sqrt{n})\, x + \theta$.

Let Θ_0 be a convex, symmetric, bounded, closed GC-set of $H(K_\mu)$ such that $\bigcup_{m \in \mathbf{N}} (u(m)\, \Theta_0)$ is dense in $H(K_\mu)$ for a strictly increasing sequence of positive numbers converging to ∞. From the characterization Theorem 2.1 for GC-sets and the properties of h, it follows that $\Theta_m = (u(m)\, \Theta_0) \cap \Theta$ constitutes a sieve of compact GC-sets in Θ. In order to point out such a set Θ_0, we now state the following lemma.

Lemma 3.1 (The sieve choice). *Let $U_{\mathbf{B}^*}$ be the unit ball of $\mathbf{B}^*$ and let $(u(m))_{m \geqslant 1}$ be any strictly increasing sequence of positive numbers converging to ∞. If $\Theta_m = (u(m)\, i^*(U_{\mathbf{B}^*})) \cap \Theta$, the sequence $(\Theta_m)_{m \geqslant 1}$ is a sieve of compact GC-sets in Θ.*

Proof. The set $i^*(U_{\mathbf{B}^*})$ is a convex, symmetric, closed, bounded set in $H(K_\mu)$. Since μ is a Radon measure, it is also a GC-set (see [5]). According to Feldman (see Section 2) it is therefore a compact GC-set. It remains to remark that $\bigcup_{m \in \mathbf{N}} (u(m)\, i^*(U_{\mathbf{B}^*}))$ is dense in $H(K_\mu)$. ∎

LEMMA 3.2. *Let L be the isonormal process on $H(K_\mu)$. If $\rho(\theta) = L(\theta) - \frac{1}{2}\|\theta\|_\mu^2$, then for any $m \geqslant 1$ and for μ-almost every x in $\mathbf{B}$ there exists a unique $\hat{\theta}_m(x)$ in Θ_m such that $\rho(\hat{\theta}_m(x))(x) = \operatorname{Sup}_{\theta \in \Theta_m} \rho(\theta)(x)$.*

Proof. For any integer $m \geqslant 1$, Θ_m equipped with the distance induced by the norm of $H(K_\mu)$ is a compact metric space. Moreover, since $H(K_\mu)$ is separable (see [15]) and Θ_m is a GC-set, the restriction of ρ to Θ_m can be viewed as a random vector defined on $(\mathbf{B}, \mathscr{B}, \mu)$ with values in $\mathbf{C}[\Theta_m]$, where $\mathbf{C}[\Theta_m]$ denotes the Banach space of real continuous functions on Θ_m with the supremum norm. Hence, none of the maximum likelihood sets M_m is empty. The lemma follows from the fact that μ-almost surely, the images of local maxima of $(\rho(\theta); \theta \in \Theta_m)$ are distinct. Indeed, for that it is sufficient to show that, for any couple of disjoint closed sets in Θ_m, say I and J:

$$\mu(\{x \in \mathbf{B}; \max_{\theta \in I} \rho(\theta)(x) = \max_{\theta \in J} \rho(\theta)(x)\}) = 0, \tag{3.2}$$

since $\{x \in \mathbf{B};$ two maximum have the same ordinate$\} \subseteq \bigcup_{I,J} \{x \in \mathbf{B}; \max_{\theta \in I} \rho(\theta)(x) = \max_{\theta \in J} \rho(\theta)(x)\}$, where the union is taken over all pairs of disjoint closed balls in Θ_m with rational radius (countable since Θ_m is separable). To prove (3.2) it is sufficient to notice that, if $0 \notin J$ then $Y = \operatorname{Max}_{\theta \in J} \rho(\theta)$ has an absolutely continuous distribution with respect to the Lebesgue measure on $\mathbf{R}$ (see Theorem 8.1 of [8] or the analog in [20]) and that $\mu(\operatorname{Max}_{\theta \in I} \rho(\theta) = Y / Y = y)$ is zero by the remark following the proof of Theorem 8.1 of [8]. This suffices to prove the lemma.

Remark 3.1. We stated Lemma 3.2 with respect to the measure μ in order to simplify the notation. It is obvious that the same result holds for any measure μ^n of (3.1).

Given the size n of the random sample, our aim now is to find an appropriate growth size $u(m)$ for the sieve $(\Theta_m)_{m \geqslant 1}$ introduced above, in order to ensure the consistency of the sequence of the maximum likelihood sieve estimators. Let $(\varepsilon(m))_{m \geqslant 1}$ be any decreasing sequence of positive numbers converging to zero and let us define the subset A_m of Θ_m by

$$A_m = \{\theta \in \Theta_m; H(\theta_0; \theta) \geqslant H(\theta_0; \Theta_m) - \varepsilon(m)\}. \tag{3.4}$$

The Kullback–Leibler information number $H(\theta_0; \theta_0) - H(\theta_0, \theta)$ is a measure of the error in approximating θ_0 by θ. The next lemma suggests as a "natural" metric for the estimation problem, the norm of the RKHS of the measure μ.

LEMMA 3.3. *The sequence of sets A_m defined by* (3.4) *is such that* $\sup_{\theta \in A_m} \|\theta - \theta_0\|_\mu \to 0$ *as* $m \to +\infty$.

Proof. A_m is a compact nonempty set of Θ_m by construction. Thus, we can chose a_m in A_m such that $\|a_m-\theta_0\|_\mu = \mathrm{Max}_{a\in A_m}\|a-\theta_0\|_\mu$. By definition (3.4) we have

$$H(\theta_0;\Theta_m)\geqslant H(\theta_0;a_m)\geqslant H(\theta_0;\Theta_m)-\varepsilon(m). \tag{3.5}$$

Now, a simple computation leads to

$$H(\theta_0;\theta_0)-H(\theta_0;\theta)=\tfrac{1}{2}\|\theta-\theta_0\|_\mu^2 \tag{3.6}$$

for any θ in Θ. Since $\bigcup_{n\in\mathbf{N}}\Theta_n$ is dense in Θ, there is a sequence $(b_m)_{m\in\mathbf{N}}$ such that b_m is in Θ_m and $H(\theta_0;b_m)\to H(\theta_0;\theta_0)$ as m tends to infinity. From (3.6) it follows that $H(\theta_0;\Theta_m)\to H(\theta_0;\theta_0)$ as $m\to\infty$ and the conclusion follows from expressions (3.5) and (3.6).

We are now able to state:

THEOREM 3.1. *If $u(m_n)=O(n^{1/2-\varepsilon})$ for some $\varepsilon>0$, then the maximum likelihood sieve estimator $\hat\theta^n_{m_n}$ of Lemma 3.2 converges almost surely to θ_0 in the metric of Θ.*

Proof. Let A_{m_n} be the set of Lemma 3.3 corresponding to $\varepsilon(m_n)=(\log n)^{-1/2}$. If $M^n_{m_n}(x)$ is the maximum likelihood set in Θ_{m_n}, by Lemma 3.2 we know that for μ^n-almost every x in $\mathbf{B}$, $M^n_{m_n}(x)=\{\hat\theta^n_{m_n}(x)\}$. Let Z_{m_n} be the subset of $\mathbf{B}$ defined by

$$Z_{m_n}=\{x\in\mathbf{B};\ \hat\theta^n_{m_n}(x)\notin A_{m_n}\}.$$

One can see easily that

$$\begin{aligned} Z_{m_n} &\subseteq \{x\in\mathbf{B};\ \sup_{\theta\in\Theta_{m_n}\setminus A_{m_n}}\rho(\theta)(x)\geqslant\sup_{\theta'\in A_{m_n}}\rho(\theta')(x)\}\\ &\subseteq \{x\in\mathbf{B};\ \sup_{\theta\in\Theta_{m_n}\setminus A_{m_n}}\rho(\theta)(x)\geqslant\rho(\theta_{m_n})(x)\},\end{aligned}$$

where θ_{m_n} is such that $H(\theta_0;\Theta_{m_n})=H(\theta_0;\theta_{m_n})$. For $y=(x-\theta_0)/\sigma_n$, where $\sigma_n=n^{-1/2}$ we have

$$\rho(\theta)(x)-\rho(\theta_{m_n})(x)=\sigma_n L(\theta-\theta_{m_n})(y)+H(\theta_0;\theta)-H(\theta_0;\theta_{m_n}).$$

For any θ in $\Theta_{m_n}\setminus A_{m_n}$, $H(\theta_0;\theta)-H(\theta_0;\theta_{m_n})<-\varepsilon(m_n)$. Hence, for such θ,

$$\rho(\theta)(x)-\rho(\theta_{m_n})(x)\leqslant\sigma_n L(\theta-\theta_{m_n})(y)-\varepsilon(m_n). \tag{3.7}$$

Let $\Psi_{m_n}(y) = \sup_{\theta \in \Theta_{m_n} \setminus A_{m_n}} L(\theta - \theta_{m_n})(y)$. The function Ψ_{m_n} belongs to the class $\mathrm{Lip}(\mu, h(\Theta_{m_n}))$ (see [14, Example 3, p. 23]). Noting that

$$\begin{aligned}\mu_{\theta_0}^n(Z_{m_n}) &\leqslant \mu_{\theta_0}^n(\{x \in \mathbf{B};\ \sup_{\theta \in \Theta_{m_n} \setminus A_{m_n}} [\rho(\theta)(x) - \rho(\theta_{m_n})(x)] \geqslant 0\}) \\ &\leqslant \mu(\{y \in \mathbf{B};\ \sup_{\theta \in \Theta_{m_n} \setminus A_{m_n}} [\sigma_n L(\theta - \theta_{m_n})(y) - \varepsilon(m_n)] \geqslant 0\}) \\ &= \mu(\{y \in \mathbf{B};\ \sigma_n \Psi_{m_n}(y) - \varepsilon(m_n) \geqslant 0\})\end{aligned}$$

and using Corollary 1 of [14, p. 26], we obtain

$$\mu_{\theta_0}^n(Z_{m_n}) \leqslant \exp\left\{-\frac{1}{2h(\Theta_{m_n})^2}\left[\frac{\varepsilon(m_n)}{\sigma_n} - \frac{1}{\sqrt{2\pi}} h(\Theta_{m_n})\right]^2\right\}.$$

But according to the properties of h (see Section 2), $h(\Theta_{m_n}) \leqslant u(m_n)\, h(\Theta_0) = O(n^{1/2-\varepsilon})$. Hence, $\mu_{\theta_0}^n(Z_{m_n}) \leqslant K \exp\{-\frac{1}{2}[n^{\varepsilon}(\log n)^{-1/2}]^2\}$. It follows from the Borel–Cantelli lemma that $\mu_{\theta_0}^{(\infty)}(Z_{m_n}\ \text{i.o.}) = 0$ and now, by Lemma 3.3, we have, $\lim_{n \to \infty} \hat{\theta}_{m_n}^n(x) = \theta_0$, $\mu_{\theta_0}^{(\infty)}$-almost surely.

PROPOSITION 3.1. *Assume the conditions of Theorem* 3.1 *hold. Then* $\lim_{n \to \infty} E_{\theta_0}(\|\hat{\theta}_{m_n}^n - \theta_0\|_\mu^2) = 0$. *Moreover, for every* $\delta > 0$ *there exists an integer* $N(\delta, \theta_0)$ *such that, for any* $n \geqslant N$, *we have*

$$\mu_{\theta_0}^n(\{x \in \mathbf{B};\ \|\hat{\theta}_{m_n}^n(x) - \theta_0\|_\mu^2 \geqslant \delta\}) \leqslant \exp(-v_n) \qquad \textit{where} \quad v_n = O(n^{2\varepsilon}).$$

Proof. By expression (3.6),

$$E_{\theta_0}(\|\hat{\theta}_{m_n}^n - \theta_0\|_\mu^2) = 2E_{\theta_0}(H(\theta_0; \theta_0) - H(\theta_0; \hat{\theta}_{m_n}^n)).$$

Now, $E_{\theta_0}(H(\theta_0; \theta_0) - H(\theta_0; \hat{\theta}_{m_n}^n)) = H(\theta_0; \theta_0) - H(\theta_0; \theta_{m_n}) + E_{\theta_0}(H(\theta_0; \theta_{m_n}) - H(\theta_0; \hat{\theta}_{m_n}^n))$, where θ_{m_n} has been introduced in the proof of the Theorem 3.1. We have then

$$\begin{aligned}&E_{\theta_0}(H(\theta_0; \theta_{m_n}) - H(\theta_0; \hat{\theta}_{m_n}^n)) \\ &\quad = \int_{\hat{\theta}_{m_n}^n \in A_{m_n}} (H(\theta_0; \theta_{m_n}) - H(\theta_0; \hat{\theta}_{m_n}^n))\, \mu_{\theta_0}^{(\infty)}(dx) \\ &\qquad + \int_{\hat{\theta}_{m_n}^n \notin A_{m_n}^c} (H(\theta_0; \theta_{m_n}) - H(\theta_0; \hat{\theta}_{m_n}^n))\, \mu_{\theta_0}^{(\infty)}(dx) \\ &\quad \leqslant \mu_{\theta_0}^{(\infty)}(\hat{\theta}_{m_n}^n \in A_{m_n})\, \varepsilon(m_n) + c\mu_{\theta_0}^{(\infty)}(\hat{\theta}_{m_n}^n \notin A_{m_n})\, u(m_n),\end{aligned}$$

where c is a generic constant (by the fact that both $\hat{\theta}_{m_n}^n$ and θ_{m_n} belong to

$\Theta_{m_n} = u(m_n)\,\Theta_0$). The first result of the proposition follows now from the bound on $\mu_{\theta_0}^{(\infty)}(\hat{\theta}^n_{m_n} \notin A_{m_n})$ in Theorem 3.1. Similarly,

$$\begin{aligned}
&\mu_{\theta_0}^{(\infty)}\{\|\hat{\theta}^n_{m_n} - \theta_0\|_\mu^2 > \delta\} \\
&\quad = \mu_{\theta_0}^{(\infty)}\left\{(H(\theta_0;\theta_0) - H(\theta_0;\hat{\theta}^n_{m_n})) > \frac{\delta}{2}\right\} \\
&\quad = \mu_{\theta_0}^{(\infty)}\left\{(H(\theta_0;\theta_0) - H(\theta_0;\theta_{m_n}) + H(\theta_0;\theta_{m_n}) - H(\theta_0;\hat{\theta}^n_{m_n})) > \frac{\delta}{2}\right\} \\
&\quad = \mu_{\theta_0}^{(\infty)}\left\{H(\theta_0;\theta_{m_n}) - H(\theta_0;\hat{\theta}^n_{m_n})) > \frac{\delta}{2} - (H(\theta_0;\theta_0) - H(\theta_0;\theta_{m_n}))\right\}.
\end{aligned}$$

By Lemma 3.3, $\lim_{n\to\infty}(H(\theta_0;\theta_0) - H(\theta_0;\theta_{m_n})) = 0$. Therefore there exists a $N(\theta_0, \delta)$ such that, for $n \geqslant N(\theta_0, \delta)$, $\delta/2 - (H(\theta_0;\theta_0) - H(\theta_0;\theta_{m_n})) > \delta' > 0$. As for the relevant part in the proof of Theorem 3.1, one still has, for $\varepsilon(m_n) = \delta'$, $\mu_{\theta_0}^{(\infty)}(Z_{m_n}) \leqslant O(\exp - (\delta'^2 n^{2\varepsilon}/2))$. Since δ is arbitrary, Proposition 3.1 holds.

Remark 3.2. The term $H(\theta_0;\theta_0) - H(\theta_0;\theta_{m_n})$ can be interpreted as a measure of our ignorance of θ_0. It represents the deterministic error when one uses a sieve to "approximate" the hole parameter set and has nothing to do with the stochastic error which has been taken care in Theorem 3.1.

A special case of Theorem 3.1 is discussed in Geman and Hwang [11], where the growth size is $u(m_n) = O(n^{1/6-\varepsilon})$. One can see that the speed given in Theorem 3.1 in a general setup is better. Another remark to be made is that $\varepsilon(m_n)/(\sigma_n u(m_n))$ always has to be greater than $(2\pi)^{-1/2} h(\Theta_0)$.

We now turn to some applications of the results of Section 3.

4. Examples

The examples are chosen for illustration. But first, let us briefly recall, in general terms, some properties about simple subsets of Hilbert spaces.

Let $\mathbf{H}$ be a separable Hilbert space, $\{\phi_n\}_{n\in\mathbf{N}}$ an orthonormal basis of $\mathbf{H}$, and let $\{b_n\}$ be a positive sequence of real numbers, which is square integrable. The subset $E(\{b_n\}, \{\phi_n\})$ of $\mathbf{H}$, defined by

$$E(\{b_n\}, \{\phi_n\}) = \left\{x = \sum_{n\in\mathbf{N}} x_n\phi_n;\ \sum_{n\geqslant 1} x_n^2/b_n^2 \leqslant 1\right\} \tag{4.1}$$

is called a Schmidt ellipsoid of $\mathbf{H}$ adapted to $\{\phi_n\}_{n\in\mathbf{N}}$ and $\{b_n\}_{n\in\mathbf{N}}$ (see [9]). As pointed out by S. Chevet [5], a Schmidt ellipsoid is a compact convex and symmetric GC-set of $\mathbf{H}$.

EXAMPLE 4.1. Suppose that we make repeated and independent observations in the statistical space

$$\{\mathbf{B}, \mathscr{B}, \{\mu_\theta; \theta \in \Theta\}\}, \tag{4.2}$$

where μ is a mean zero Gaussian measure on $\mathscr{B}$ and Θ is an infinite dimensional subspace of the RKHS $H(K_\mu)$ of μ for which there is a subset $\{e_j^*; j \geq 1\}$ of $\mathbf{B}^*$ such that $\{\phi_j = i^*(e_j^*)\}$ is a complete orthonormal system in Θ. Let $E(\{b_n\}, \{\phi_n\})$ be a Schmidt ellipsoid of $H(K_\mu)$ and set $\Theta_0 = E(\{b_n\}, \{\phi_n\})$. Hence Θ_0 is a compact convex symmetric GC-set of Θ and, according to the results of Section 3, $(\Theta_m)_{m \geq 1}$, where $\Theta_m = u(m)\,\Theta_0$ with $u(m)$ an increasing sequence of positive numbers converging to infinity, is a sieve of compact GC-sets of Θ.

A direct application of Theorem 3.1 establishes the following theorem.

THEOREM 4.1. *In Θ_m, the maximum likelihood solution given n i.i.d. observations $x^1, x^2, \ldots, x^n$ in the statistical space $\{\mathbf{B}, \mathscr{B}, \{\mu_\theta; \theta \in \Theta\}\}$ is*

$$\hat{\theta}_m^n(x^1, x^2, \ldots, x^n) = \sum_{k \geq 1} \hat{\theta}_k \phi_k,$$

where

$$\hat{\theta}_k = \frac{\langle e_k^*, \bar{x}_n \rangle}{\{1 + \lambda_m / b_k^2\}}, \quad \sum_{k \geq 1} \frac{\hat{\theta}_k^2}{b_k^2} = u(m)^2, \qquad \bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x^i.$$

If $m_n \to \infty$ in such a way that $u(m_n) = O(n^{1/2 - \varepsilon})$ for some $\varepsilon > 0$, then

$$\hat{\theta}_{m_n}^n \to \theta_0 \qquad \mu_{\theta_0}^{(\infty)}\text{-almost surely.}$$

Remark 4.1. A particular case of the above example is discussed in [11] where consistency in the sense of our Theorem 3.1 is proven. In their example, $\Theta = H(K_\mu)$ and Θ_m is the closed ball of radius m in $H(K_\mu)$. It is easy to see that $\Theta_m = m^{1/2}\Theta_0$, where Θ_0 is the closed unit ball of $H(K_\mu)$ which is the variance ellipsoid of μ (see [18, p. 10]). Their approach leads to a more restricted rate of growth for m_n, namely $u(m_n) = O(n^{1/6 - \varepsilon})$.

Proof of Theorem 4.1. Our aim is to maximize

$$\frac{d\mu_\theta^{(n)}}{d\mu^{(n)}} = \exp\left\{\sqrt{n}\left[\sum_{k \geq 1}\left(\langle \theta, \phi_k \rangle_\mu \cdot \langle e_k^*, \bar{x}_n \rangle - \frac{1}{2}\langle \theta, \phi_k \rangle_\mu^2\right)\right]\right\}$$

with the constraint

$$\sum_{k \geq 1} \frac{\langle \theta, \phi_k \rangle_\mu^2}{b_k^2} = u(m)^2. \tag{4.3}$$

By the Lagrange multiplier method we get the restricted maximum likelihood solution $\hat{\theta}_k = \langle e_k^*, \bar{x}_n \rangle / \{1 + \lambda / b_k^2\}$ where λ should be adjusted to satisfy $\sum_{k \geq 1} \hat{\theta}_k^2 / b_k^2 = u(m)^2$. Note that this last sum converges a.s. for any $\lambda > 0$ and that it is a strictly decreasing function in λ, taking the value ∞ at $\lambda = 0$ and tending to 0 as $\lambda \to \infty$. Thus, there exists a.s. a unique root $\lambda(m)$ such that (4.3) is satisfied. The theorem follows by a direct application of Theorem 3.1.

For illustration of Theorem 4.1 consider the stochastic process

$$x(t) = m(t) + W_t, \qquad m \in \Theta, \tag{4.4}$$

where W is the Wiener process over $[0, T]$ with unit variance per unit time, and where Θ is the subset of H_W of periodic functions with smallest period T/q where q is a fixed integer greater than one. It is known (e.g., [2]) that

$$H_W = \left\{ f \in \mathbf{C}_0([0, T]); f(t) = \int_0^t g(u)\, du,\ g \in L^2([0, T], dt) \right\}$$

and

$$\langle f, g \rangle_W = \int_0^T \dot{f}(u)\, \dot{g}(u)\, du.$$

In order to apply Theorem 4.1, it is sufficient to prove that Θ is a closed subspace of H_W with a CONS in $\mathbf{B}^*$. The subspace Θ is closed since the convergence in H_W implies point convergence, and that any point limit of periodic functions of period T/q is still of period T/q. A basis for Θ is obtained by the system

$$S = \{u_{2kq}, k \geq 0; u_{2kq-1}, k \geq 1\},$$

where $u_0(t) = t$, $u_{2k}(t) = \int_0^t \cos((2\pi k/T)\, s)\, ds$ and $u_{2k-1}(t) = \int_0^t \sin((2\pi k/T)\, s)\, ds$ for $k \geq 1$. It is also easy to see that S is a subset of $i^*(\mathbf{B}^*)$, where $\mathbf{B} = \mathbf{C}_0([0, T])$.

EXAMPLE 4.2. Let us consider again the statistical space corresponding to the model (4.4) with $T = 1$ and $\Theta = \text{Lip } \alpha$, where

$$\text{Lip } \alpha = \{ f \in \mathbf{C}([0, 1]); f(0) = 0 \text{ and } w_f(\delta) = O(\delta^\alpha) \}$$

with $w_f(\delta) = \sup_{|x-y| < \delta} |f(x) - f(y)|$ and $\alpha > \frac{3}{2}$.

Let Θ_0 be the subset of Θ of functions f with $w_f(\delta) = \delta^\alpha$. For any increasing sequence of positive real numbers $(u(m))_{m \geq 1}$, it is not hard to

see that $\Theta_m = u(m)\,\Theta_0$ is closed, and that the union of the increasing sequence $(\Theta_m)_{m\geqslant 1}$ is dense in Θ. Hence, in order to apply Theorem 3.1, it is sufficient to check that Θ_0 is a compact GC-set of H_W. In this regard, let $(h_p)_{p\geqslant 0}$ be the orthonormal Haar system of $L^2([0,1], dt)$ (see [1]) and let $(H_p)_{p\geqslant 0}$ be the corresponding Schauder basis of $C([0,1])$, that is,

$$H_p(x) = \int_0^x h_p(u)\,du.$$

Quite obviously $(H_p)_{p\geqslant 0}$ is a CONS of H_W. For any f in H_W let $(\xi_p(f))_{p\geqslant 0}$ be the components of f on $(H_p)_{p\geqslant 0}$. We then have

$$f\in\Theta_0 \Rightarrow |\xi_p(f)|\leqslant 2/p^{\alpha-1/2}, \qquad \xi_0(f)=0.$$

Indeed, we have for any f in $\mathbf{C}_0([0,1])$ (see [1])

$$\xi_0(f) = f(1)$$

and

$$\text{for } p=2^n+k, \quad \xi_p(f) = 2^{n/2}\left[2f\left(\left[k+\frac{1}{2}\right]2^{-n}\right) - f([k+1]\,2^{-n})\right]. \tag{4.5}$$

Now, since $\alpha>1$, $f(1)=0$ for any f in Θ_0. Moreover, for any $2^n<p<2^{n+1}$, by (4.5) we have

$$|\xi_p(f)|\leqslant 2\cdot 2^{n/2}\cdot w_f(2^{-(n+1)}). \tag{4.6}$$

Setting $\delta = 2^{-(n+1)}$ in (4.6), we obtain

$$|\xi_p(f)|\leqslant 2\cdot\sqrt{p}\cdot\delta^\alpha \leqslant 2\cdot\sqrt{p}\cdot\frac{1}{p^\alpha} = \frac{2}{p^{\alpha-1/2}}, \tag{4.7}$$

since $p<2^{n+1}$. From (4.7) it follows that

$$\Theta_0\subseteq \exists_1 = \left\{f\in H_W;\ |\xi_p(f)|\leqslant\frac{2}{p^{\alpha-1/2}},\ p\geqslant 1,\ \xi_0(f)=0\right\}.$$

But $\exists_1$ is a compact subset of H_W such that $h(\exists_1) = 4\sum_{p\geqslant 1} 1/p^{\alpha-1/2} < \infty$ since $\alpha>\frac{3}{2}$ (see Corollary VIII.2.1 of [5]). Thence, Θ_0 is a GB-set and that implies that it is a compact GC-set according to Corollary VIII.6.2 of [5]. By Theorem 3.1, the above sieve makes ML estimation consistent, but in this case one should find an awkward procedure for computing the maximum likelihood solution.

5. General Comments

We have constructed a point estimator of the mean of a Gaussian process of known covariance by the method of sieves. The estimator is strongly consistent in the norm of the reproducing kernel Hilbert space of the process, but we assume that the covariance is known. When there is a partial information about the covariance operator of the Gaussian process, the method of sieves has been applied recently, by the author and J. Beder (see [3]), for a simultaneous estimation of the mean and the covariance, while, when no prior information on the covariance is available, the empirical covariance operator estimates it consistently.

We have used the facts that the MLE for the mean from n i.i.d. samples is the same as the MLE from one sample from the sufficient statistic for the mean, and the relation between the norm of the parameter space and the Kullback–Leibler information. These facts are common to infinite dimensional exponential statistical spaces introduced in [17] by J. L. Soler and it may be hoped that the previous results can be generalized to such families.

Since this paper was first written, maximum likelihood estimation of the mean of a Gaussian process with known covariance by the method of sieves has been also considered by Beder [7], but his method deals only with maximizing the likelihood over subspaces. In this work, we are dealing with general subsets of the RKHS, but we have paid a price for our generality, namely, the maximization procedure is sometimes awkward as pointed out by Example 4.2.

Acknowledgments

The author thanks a referee for his detailed comments and constructive suggestions. Thanks also go to the Department of Mathematics at the University of California, Irvine, for its financial support.

References

[1] Alexits, G. (1961). *Convergence Problems of Orthogonal Series.* Pergamon, New York.

[2] Antoniadis, A. (1984). Analysis of variance on function spaces. *Math. Oper. Statist. Ser. Statist.* **15**, No. 1, 59–71.

[3] Antoniadis, A., and Beder, J. H. (1986). Joint estimation of the mean and the covariance of a Banach space-valued Gaussian vector. Submitted.

[4] Azencott, R. (1980). Grandes deviations et applications. *In* Lecture Notes in Mathematics Vol. 774. Springer-Verlag, Berlin/New York.

[5] Badrikian, A., and Chevet, S. (1974). *Mesures Cylindriques, Espaces de Wiener et Fonctions Aleatoires Gaussiennes.* Lecture Notes in Mathematics Vol. 379. Springer-Verlag, New York.

[6] BAHADUR, R. R., AND ZABELL, S. L. (1979). Large deviations of the sample mean in general vector spaces. *Ann. Probab.* **7** 587–621.

[7] BEDER, J. H. (1987). A sieve estimator for the mean of a Gaussian process. *Ann. Statist.* **15**, No. 1, 59–78.

[8] DUDLEY, R. M. (1973). Sample functions of the Gaussian process. *Ann. Probab.* **1**, No. 1, 66–103.

[9] DUDLEY, R. M. (1967). The sizes of compact subsets of Hilbert spaces and continuity of Gaussian processes. *J. Funct. Anal.* **1** 290–330.

[10] FELDMAN, J. (1971). Sets of boundedness and continuity for the canonical normal process. In *Proceedings, Sixth Berkeley Sympos. Math. Statist. Probab., Vol. 2,* pp. 357–368.

[11] GEMAN, S., AND HWANG, C. R. (1982). Non parametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10**, No. 2, 401–414.

[12] GRENANDER, U. (1981). *Abstract Inference.* Wiley, New York.

[13] GROSS, L. (1965). Abstract Wiener spaces. In *Proceedings, Fifth Berkeley Sympos. Math. Statist. Probab.*, pp. 31–41.

[14] IBRAGIMOV, I. A., SUDAKOV, V. N., AND TSIREL'SON, B. S. (1976). Norms of Gaussian sample functions. *In* Lecture Notes in Mathematics Vol. 550, pp. 20–41. Springer-Verlag, New York.

[15] KUELBS, J. (1970). Gaussian measures on a Banach space. *J. Funct. Anal.* **5** 354–367.

[16] MCKEAGUE, I. (1985). *The method of sieves: A survey of recent applications.* Statistics Report M-718, Florida State University, Tallahassee; *Encyclopedia of Statistical Sciences.* Wiley, New York, in press.

[17] SOLER, J. L. (1977). Infinite dimensional exponential type statistical spaces (generalized exponential families). In *Recent Developments in Statistics* (J. R. Barra, Ed.), pp. 269–284. North-Holland, New York.

[18] SUDAKOV, V. N. (1979). Geometrical problems in the theory of infinite dimensional probability distributions. *Proc., Steklov Inst. Math.* **2.**

[19] VAKHANIA, N. N. (1981). *Probability Distributions on Linear Spaces.* North-Holland, New York.

[20] YLVISAKER, N. D. (1968). A note on the absence of tagencies in Gaussian sample paths. *Ann. Math. Statist.* **39** 261–262.